Length minimizing property of great circles

For definiteness, we shall work with the standard unit sphere in $\mathbf{R}^3$ defined by the equation $x^2 + y^2 + z^2 = 1$. A **great circle** on such a sphere is a circle whose center is equal to the center of the sphere itself (in this case, the origin). A fundamental result in geometry states that the shortest curve on a sphere joining two points is given by a piece of a great circle, and the goal of this document is to provide a relatively elf-contained proof of this fact using the machinery developed in Unit III of the course. At a few points it will be necessary to use material from more advanced courses on functions of a real variable; these will be noted as they appear.

Definitions and preliminaries

If we are given a great circle on a sphere and it lies on a plane $\mathbf{P}$, then either unit normal vector $\mathbf{b}$ to $\mathbf{P}$ is perpendicular to $\mathbf{P}$, and therefore one may parametrize such a curve by arc length using a formula of the form

$$\gamma(\theta) \quad - \quad (\cos\theta)\,\mathbf{a} \ + \ (\sin\theta)\,\big(\mathbf{b}\times\mathbf{a}\big)$$

where $\mathbf{a}$ is a unit vector in the plane $\mathbf{P}$

The unit sphere is an extremely symmetric object, and in particular, if $\mathbf{u}$ is an arbitrary point of the sphere, then there is a smooth isometry φ from the sphere to itself sending the "North Pole" point $\mathbf{e}_3$ to $\mathbf{u}$. This isometry is given by $\varphi(\mathbf{v}) = A\mathbf{v}$, where A is a 3×3 orthogonal matrix, and the linearity properties of A imply that under this mapping great circles through the north pole are sent to great circles through its image. Furthermore, the length of a piecewise smooth curve α on the unit sphere will be equal to the length of its image $\varphi\circ\alpha$. We shall use the existence of the specified isometries to simplify our work as follows:

INITIAL REDUCTION. *It suffices to prove the length minimizing property for piecewise smooth curves starting at the north pole.*

This is true because the isometry sends piecewise smooth curves starting at the north pole to piecewise smooth curves starting at the image point $\mathbf{u}$, and the length of the image curve is equal to the length of the original curve.■

The right choice of parametrization will also be very helpful, and it is given as follows:

$$\mathbf{X}(r,\theta) \quad = \quad \left(r\cos\theta,\, r\sin\theta,\, \sqrt{1-r^2}\,\right)$$

With respect to this parametrization the First Fundamental Form reduces to $dr\,dr \ + \ r^2\,d\theta\,d\theta$. This expression is simple, but it has one logical difficulty; strictly speaking it does not work when $r = 0$.

At the beginning of Unit III we mentioned that one can often ignore such problems when working with specific examples, and in fact this is possible here. There are two main reasons for this. First of all, if we have a piecewise smooth curve γ whose length we wish to compute using the length formula and the curve is defined on the closed interval $[a,b]$, then we may view the length as an improper integral

$$\int_a^b \sqrt{(r')^2 \ + \ r^2(\theta')^2}\,dt$$

which is the limit of integration from $a - \varepsilon$ to b as $\varepsilon \to 0$ [For the sake of completeness, here is the proof: We know that the curve has a positive length and the partial integrals are less than or equal to the entire length, and we also know that the integral from $a - \varepsilon$ to b increases as ε decreases to zero, so that the improper integral actually has a limit, which is the least upper bound of the partial integrals; continuity considerations then imply that the limit must be the length of the curve].

The preceding observation is useful if we have a curve γ such that a is the only parameter value such that $\gamma(t)$ is equal to the north pole, for then we can use polar coordinates to parametrize the curve on all of the half open interval $(a, b]$. Therefore we would like to know that we can restrict attention to such curves. This requires two steps. The first is somewhat sophisticated but should be intuitively clear; the proof will require a fairly deep understanding of continuity and the real numbers as in a first real variables course.

PROPOSITION. *If γ is a continuous curve from a closed interval $[a, b]$ to $\mathbf{R}^n$, then there is is a unique parameter value $a^* < b$ such that $\gamma(a^*) = \gamma(a)$ and a^* is maximal with respect to this property.*

Sketch of proof. (SOMEWHAT ADVANCED) Let a^* be the least upper bound of all points t on the interval (a bounded set) such that $\gamma(t) = \gamma(a)$. The set in question is nonempty because it contains a, and since it lies inside the original interval, we must have $a^* \in [a, b]$. If $\gamma(a^*) = \gamma(a)$, then a^* has all the required properties. There are two cases. If $a^* = a$, then the conclusion is trivially true. On the other hand, if $a^* > a$, the the least upper bound property implies there is a sequence of points $t_n \leq a^*$ such that $a^* - t_n > \frac{1}{n}$ and $\gamma(t_n) = \gamma(a)$. It follows that $a^* = \lim_{n\to\infty} t_n$ and by the continuity of γ we must then have

$$\gamma(a^*) \quad = \quad \lim_{n\to\infty} \gamma(t_n) \quad = \quad \lim_{n\to\infty} \gamma(a) \quad = \quad \gamma(a)$$

as required.■

The next observation is more elementary.

PROPOSITION, *If in the preceding result the curve γ is piecewise smooth and regular, then the length of γ from a to b is greater than or equal to the length from a^* to b, with equality if and only if $a = a^*$.*

Before proving this, we show how it leads to our next simplification.

SECOND REDUCTION. *It suffices to prove the length minimizing property for piecewise smooth curves γ starting at the north pole such that $\gamma(t)$ is equal to the north pole only when t is the left hand end point.*

This is true because if we have a curve that does not satisfy the given property and we restrict to $[a^*, b]$, then the restricted curve is still of the right type and its length is strictly less than the length of the original curve.■

Proof of Proposition. The integrand in the arc length formula is nonnegative by definition, and it is positive except for at most finitely many points. Therefore if $a^* > a$, then the restricted integral will be strictly smaller than the original one.■

Statement of results and proof for short curves

Here is the first main result:

WEAK LENGTH MINIMIZING PROPERTY. *Let γ be a piecewise smooth curve on the sphere which joins the north pole to some other point, and assume that the length of γ is minimal among all such curves. Then the image of γ lies on a great circle.*

Since our parametrization for the sphere only covers points that are strictly to the north of the equator, we need to begin by restricting attention to curves whose images lie in this subset. Such a curve will be called a SHORT CURVE. More generally, we shall use this term to denote an arbitrary curve starting at some point $\mathbf{w}$ on the sphere whose image is contained in the half space $\mathbf{w} \cdot \mathbf{x} > 0$; *i.e.*, it lies on one side of the great circle defined by $|\mathbf{x}|^2 = 1$ and plane through the origin defined by $\mathbf{x} \cdot \mathbf{w} = 0$ (the great circle for which $\mathbf{w}$ is a polar point). The isometries of a sphere defined by orthogonal matrices send short curves starting at one point to short curves starting at its image.

Proof of weak length minimizing for short curves. By previous observations, we may restrict attention to curves satisfying the condition in the second reduction stated above.

We may then write our short curve in polar coordinates as

$$\gamma(t) \quad = \quad \left(r(t)\cos\theta(t),\, r(t)\sin\theta(t),\, \sqrt{1 - r(t)^2}\, \right)$$

where r and θ are piecewise smooth functions and $\lim_{t \to a}\ r(t) = 0$ (for small pieces of the curve this is straightforward to do; it is less trivial for larger pieces and actually requires results in first year graduate topology courses, so here we shall simply say it can be done). Suppose we define a new curve β by replacing the function $\theta(t)$ with the constant $\theta(b)$. This curve joins the same two points, and we would like to compare the lengths L_γ and L_β. The previously derived formula implies that

$$L_\gamma \quad = \quad \int_a^b \sqrt{(r')^2 \,+\, r^2(\theta')^2}\, dt$$

and also that

$$L_\beta \quad = \quad \int_a^b \sqrt{(r')^2}\, dt \quad = \quad \int_a^b |r'|\, dt\ .$$

Now the second integral is less than or equal to the first, and the image of the second curve lies on the great circle containing the north pole and the point $\gamma(b) = \beta(b)$. Therefore we know that for each short curve joining two points there is a curve on a great circle joining two points whose length is less than or equal to the original length. Suppose now that the original curve does not lie on a great circle. This means that the function θ is not constant, so its derivative is nonzero somewhere. By continuity, it must be nowhere zero on a small interval $[c,d]$, and over this interval the length integral for γ must be strictly larger than the length integral for β. But this means than one has a similar strict inequality for the integrals over the entire intervals as well.■

The preceding result shows that curves of minimum length lie on great circles; however, experience suggests that of all such curves, a minor great circle arc should have minimum length. Here is a statement of the result.

STRONG LENGTH MINIMIZING PROPERTY FOR SHORT CURVES. *Let γ be a piecewise smooth curve on the sphere which joins the north pole to some other point, and assume that the length of γ is minimal among all such curves. Then γ is a reparametrization of the minor arc of a great circle.*

Sketch of proof. First of all, the short curve hypothesis, the second reduction and continuity imply that the image of the curve lies on that piece of the great circle above the equator; in other words, it lies on a parametrized curve

$$\Big(\cos u \cdot \cos\theta_0,\ \cos u \cdot \sin\theta_0,\ \sin u \Big)$$

where

$$-\frac{\pi}{2} < u < \frac{\pi}{2}.$$

Thus a short curve which lies on such a great circle and satisfies the condition in the second reduction will have this form where $u = f(t)$ is a piecewise smooth function such that $f(a) = 0$ and $f(b)$ is strictly between 0 and $\pi/2$. By the length formula, its length is equal to the integral of $|f'|$ over this interval; the latter is not continuous on the interval, but it only has finitely many discontinuities and can be integrated fairly directly.

Suppose now that the curve in question has minimum length. One can now apply Exercise $2(a)$ in the Additional Exercises for Section I.3 (see `dgexercises2006.*`) to conclude that f must be positive everywhere that it is defined, and from this one concludes that the given curve is a reparametrization of a great circle's minor arc.■

Generalization to arbitrary curves

One key step in the process is the following decomposition property, which requires results from real variables:

DECOMPOSITION INTO SHORT CURVES. *Given a piecewise smooth curve γ on the sphere defined on an interval $[a, b]$, it is possible to partition the interval into subintervals at points*

$$a = t_0 < t_1 < \cdots < t_M = b$$

such that the restriction of γ to each subinterval $[t_{i-1}, t_i]$ is a short curve.

This is an immediate consequence of a property called uniform continuity; we shall not attempt to explain the details here.■

We shall now indicate how one can use the decomposition to prove that an arbitrary curve of minimal length must be a great circle. If γ has minimal length, then for each i the restriction of γ to the subinterval $[t_{i-1}, t_i]$ must be a curve of minimal length joining its endpoints; if it were not, then one could replace γ over that interval by a shorter curve. This would yield a new curve $\gamma^\#$ over the entire interval joining the original two points, and its length would be strictly less than the curve of minimal length. Thus the restricted curves must also have the minimal length property and by the previous results for short curves each one must be a minor arc of a great circle. We need to show these curves are all minor arcs on the **same** great circle.

The following elementary fact from spherical geometry will be helpful:

LEMMA. *If $\mathbf{y}$ and $\mathbf{z}$ are points on the sphere such that the distance between them is less than $\sqrt{2}$, then both points lie on the same hemisphere determined by the great circle on the plane with equation $\mathbf{y} \cdot \mathbf{x} = 0$.*

Proof. Algebraically, the conclusion translates into the inequality $\mathbf{y} \cdot \mathbf{z} > 0$. The distance condition may be restated as

$$|\mathbf{y} - \mathbf{z}| < \sqrt{2}$$

and if we square this, expand $|\mathbf{y} - \mathbf{z}|^2 = (\mathbf{y} - \mathbf{z}) \cdot (\mathbf{y} - \mathbf{z}) = |\mathbf{y}|^2 - 2(\mathbf{y} \cdot \mathbf{z}) + |\mathbf{z}|^2$, and recall that $|\mathbf{y}| = |\mathbf{z}| = 1$, we obtain the inequality

$$2 - 2(\mathbf{y} \cdot \mathbf{z}) < 0$$

which immediately translates into the desired inequality $\mathbf{y} \cdot \mathbf{z} > 0$.■

We shall now proceed to show that the restrictions of γ to $[t_{i-1}, t_i]$ and $[t_i, t_{i+1}]$ are consecutive parts of the same great circle. By continuity there are points p and q such that

$$t_{i-1} \quad < \quad p \quad < \quad t_i \quad < \quad q \quad < t_{i+1}$$

such that all points in the images $\gamma([p, t_i])$ and $\gamma([t_i, q])$ have distances less than $\frac{1}{2}$ from $\gamma(t_i)$. It follows that all points in the image $\gamma([p, q])$ have distance less than 1 from $\gamma(p)$ and hence the restriction of γ to $[p, q]$ is a short curve.

Therefore the restriction of γ to $[p, q]$ must also be a reparametrization of a great circle. But this means that the restrictions to γ to $[t_{i-1}, t_i]$ and $[t_i, t_{i+1}]$ are reparametrizations of consecutive parts of the same great circle as required. If we do this for each i between 1 and $M - 1$, we see that the short curves fit together to form a reparametrization of a great circle, and since this curve is supposed to minimize length, the great circle must be a minor arc unless the image point $\gamma(b)$ is the antipodal point (or negative) of $\gamma(a)$, in which case both great circle arcs joining the points have equal (minimum) length.∎

Generalizations to other surfaces

If γ is a great circle curve on the standard unit sphere, then γ'' is a negative multiple of γ, and thus for each parameter value t the vector $\gamma''(t)$ is perpendicular to the space of tangent vectors to the surface at $\gamma(t)$. This property plays an important role in the study of curves on an arbitrary surface with minimal length.

Definition. If Σ is a surface with orientation $\mathbf{N}$, then a smooth curve γ in Σ is a *geodesic* if for each t the second derivative $\gamma''(t)$ is a multiple of $\mathbf{N}(\gamma(t))$.

We know that curves of least length in a plane are geodesics because their second derivatives vanish, and the preceding discussion shows that curves of least length on a sphere are also geodesics. However, there are also geodesics on a sphere that do not minimize length. In particular if $\mathbf{p}$ and $\mathbf{q}$ are points on the sphere that are not antipodal, then the major arc of the great circle joining these points is a geodesic but does not minimize distances.

The precise connection between geodesics and curves of least length is reflected by the following basic fact: *If Σ is a surface and $\mathbf{p} \in \Sigma$, then there is some distance $r > 0$ such that every point $\mathbf{x}$ on Σ satisfying $|\mathbf{x} - \mathbf{p}| < r$ can be joined to $\mathbf{p}$ by a unique geodesic, and the length of this geodesic is minimal among all piecewise smooth curves joining the given two points.*

Further information on this may be found on pages 239–240 of Lipschultz as well as in DO CARMO and O'NEILL.

If we are given two points on a surface, it is not necessarily true that one can join them by a curve of minimum length; for example, if we take the surface to be the xy-plane with the origin removed, then the greatest lower bound on the lengths of curves joining $(\pm 1, 0, 0)$ is 2, but there is no curve lying entirely in this surface that joins the two curves and has length equal to 2. On the other hand, if one has a surface that is closed (under taking limits of sequences) and bounded, then one can prove that there is always a curve of minimal length joining two points, and in fact this curve is a geodesic. However, a proof of this fact is beyond the scope of this course.